TS-ICL: A Flexible Time-Indexed Foundation Model for Time Series via In-Context Learning

NeurIPS 2026

December 2026

Etienne Le Naour — EDF R&D
Tahar Nabil — EDF R&D
Adrien Petralia — EDF R&D

Introduction

Tabular Foundation Models can solve Time Series tasks

Tabular FMs as free TSFMs?

- **TabPFN-v2** [8], **TabICL-v2** [13] → powerful probabilistic tabular regressors
- Cast **forecasting as tabular regression** via temporal featurization [9]
- Natively accomodate **irregular/incomplete measurements**
- **Pretraining: synthetic tabular data only**, strictly no time series

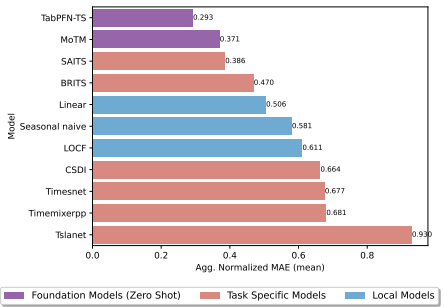


Tabular Foundation Models can solve Time Series tasks

Tabular FMs as free TSFMs?

- **TabPFN-v2** [8], **TabICL-v2** [13] → powerful probabilistic tabular regressors
- Cast **forecasting as tabular regression** via temporal featurization [9]
- Natively accomodate **irregular/incomplete measurements**
- **Pretraining: synthetic tabular data only**, strictly no time series

→ Competitive forecasting, SOTA imputation



TabPFN v2 [8] outperforms all specialised time series imputation models on `fm-impute-bench` *[12]*

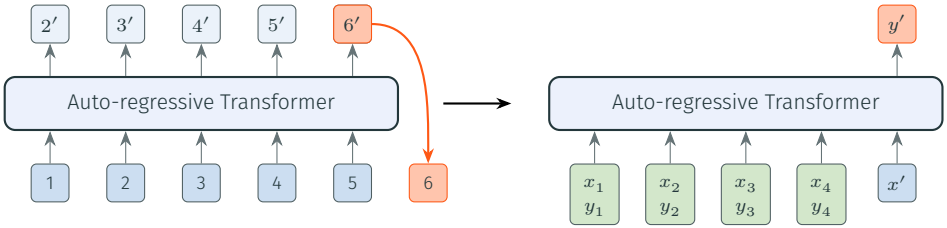
Background: Transformers as General-Purpose In-Context Learners

- A **supervised learning algorithm** is a mapping *(training set, query)* $\rightarrow$ *label*

$$(\{x_i, y_i\}_{i=1}^{N}, x') \mapsto y'$$

  that improves its predictions y' as $N \nearrow$.
- **Meta-learning** a learning algorithm: finding these function via meta-optimization - e.g. MAML [5], CAVIA [16], etc.
- Recent works connect **learning Transformer weights** with meta-learning [6, 11, 15]
- **TabPFN, TabICL**, etc. build on this paradigm for **tabular data** [7, 8, 13]

Background: Transformers as General-Purpose In-Context Learners

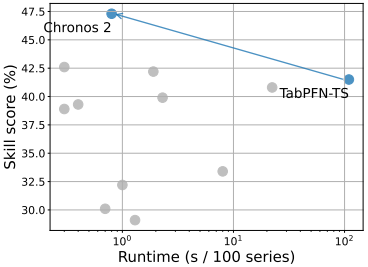


A **single forward pass** of an attention-based architecture **performs both training and inference**

Keeping *time-index* formulation, adding a *temporal* inductive bias to TFMs

Where to improve Tabular Foundation Models?

- **Bridging the performance gap** vs leading patch-based forecasting TSFMs
- **Efficient** zero-shot performance



fev-bench leaderboard (Spring 26) [14]

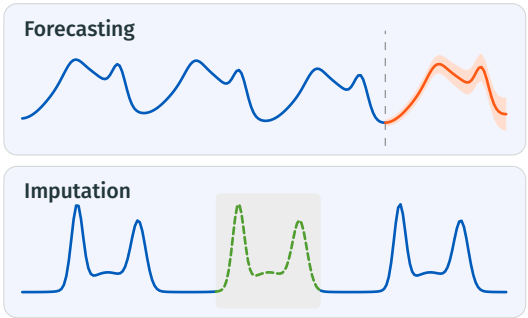
Can we **learn flexible temporal representations** that generalize across time series tasks and observation settings in **a single in-context learning framework**?

Motivation #1: a flexible architecture that unifies forecasting & imputation

Incomplete real-world measurements call for frameworks that can jointly handle forecasting and imputation

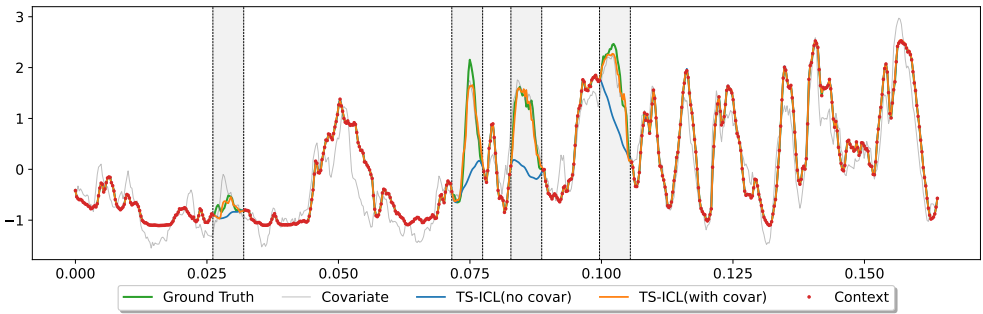
Scenarios we meet in practice:

- Irregularly-sampled / partially observed data
- Sensor completion / data quality
- Counterfactual estimation
- etc.



Motivation #2: covariate-aware inference

Native covariate-aware inference is key to unlock practical use cases



Recovering four one-day blocks of hourly wind power production data, with TS-ICL.

TS-ICL: bridging TFMs & TSFMs capabilities

With **TS-ICL**, we introduce a unified probabilistic 28M-parameter Transformer foundation model for imputation and forecasting

Table 1: Capabilities of recent time series foundation models. Only *TS-ICL* supports forecasting and imputation, while enabling efficient covariate-aware inference and supporting irregular sampling.

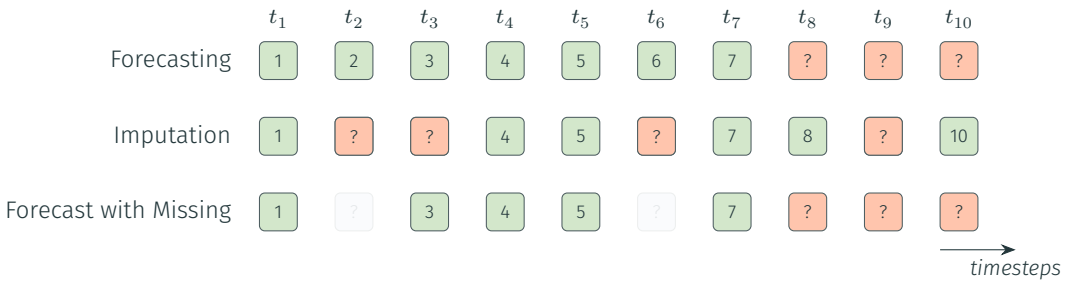
Method	Handles Forecasting	Handles Imputation	Covariate Integration	Probabilistic Predictions	Designed for Time Series	Irregular Sampling	Fast Inference
TiRex [2], *Toto* [3], *TimesFMv2.5* [4]	✓	✗	✗	✓	✓	✗	✓
Chronos-2 [1]	✓	✗	✓	✓	✓	✗	✓
TabPFNv2.5-TS [9], *TabICLv2-TS* [13]	✓	✓	✓	✓	✗	✓	✗
TS-ICL (ours)	✓	✓	✓	✓	✓	✓	✓

TS-ICL

Time series modelling as in-context regression

In-Context Learning Framework (univariate setting)

1 time series = 1 dataset = collection of observed features (timesteps t) and labels (target values $x(t)$)



Learn θ s.t. $f_\theta(\mathbf{t'}, \{\mathbf{t_i}, \mathbf{x(t_i)}\}) \approx \mathbf{x(t')}$?

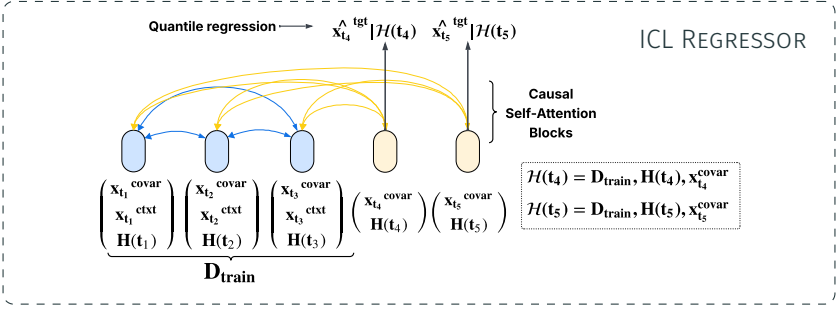
(1) Architecture overview: an encoder-regressor network

A Transformer solved the regression tasks...

- **Goal:** learn a regressor $t \mapsto x(t)$ (pointwise formulation)

→ **ICL Transformer**, learns to map query *features* to *labels*, cond. on observed *(query, label)* pairs

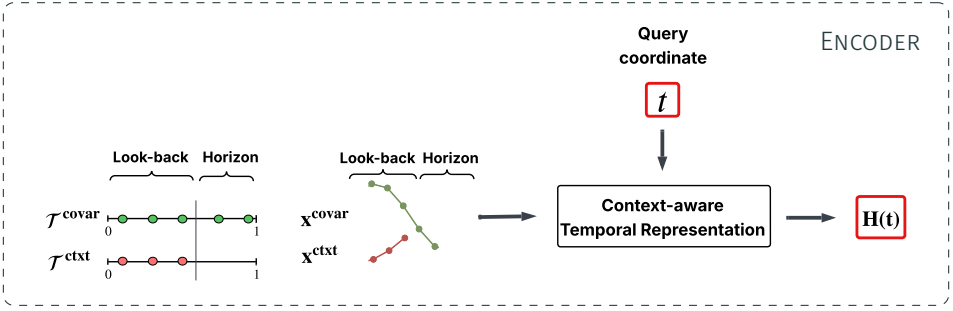
- Features = learned encodings of time $H(t)$ / labels = target time series values



(1) Architecture overview: an encoder-regressor network

... BASED ON LEARNED CONTEXTUAL ENCODINGS OF TIME

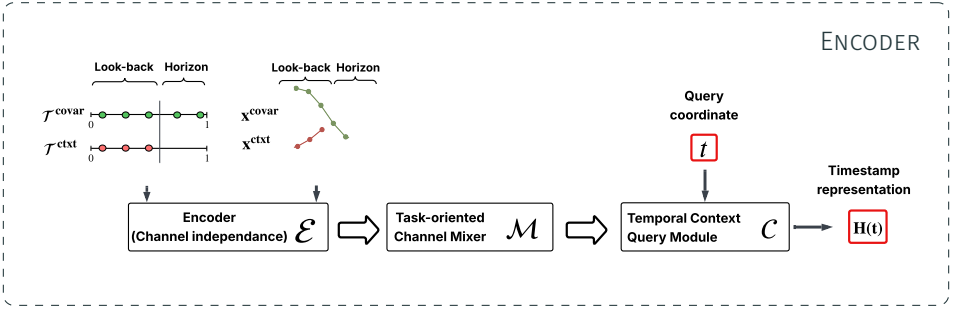
For each t, $H(t)$ is informed by **local** (rich Fourier Features) & **global context** (a compression of the available context in entirety through a Perceiver-based encoder) $\rightarrow$ heteregeneous contexts mapped to fixed dim. $H(t)$



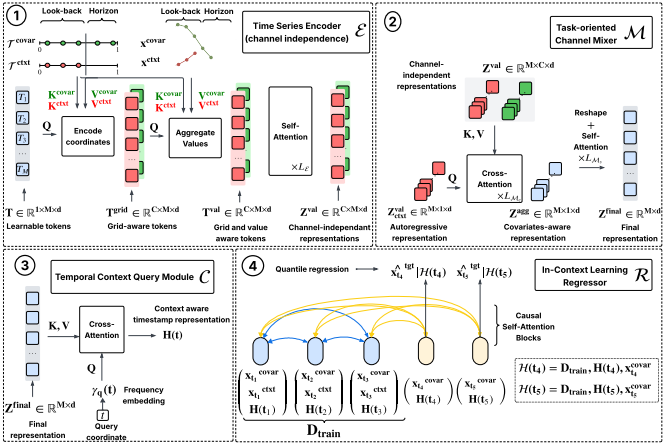
(1) Architecture overview: an encoder-regressor network

... BASED ON LEARNED CONTEXTUAL ENCODINGS OF TIME

For each t, $H(t)$ is informed by **local** (rich Fourier Features) & **global context** (a compression of the available context in entirety through a Perceiver-based encoder) $\rightarrow$ heteregeneous contexts mapped to fixed dim. $H(t)$



(1) Architecture breakdown



TS-ICL leverages the flexibility of cross-attention mechanisms to handle heterogeneous contexts

(2) Data Prior, stage 1: univariate pretraining corpus

- **Large-scale pretraining** distribution over **univariate time series**
- Comprising **40 datasets** and ≈ **2M time series** with lengths ranging from 100 to 600k time steps
- Strict **no-overlapping** with all benchmarks used for evaluation

Dataset	Release Platform	Domain	Freq	Num. Series	Num. Variates	Max Length	Weight
Australian Electricity	Chronos	Energy	30T	5	1	232,272	220
BDG-2 Bull	LOTSA	Energy	H	41	1	12,280	25
BDG-2 Fox	LOTSA	Energy	H	135	1	12,280	5
BDG-2 Panther	LOTSA	Energy	H	105	1	6,132	2.5
BuildingsBench900k	LOTSA	Energy	H	100,000†	1	8,759	0.02048
Residential Load Power	LOTSA	Energy	1T	271	3	614,880	1.2
Residential PV Power	LOTSA	Energy	1T	233	3	614,880	1.5
Wind Farms H	Chronos	Energy	H	337	1	6,148	4
Wind Farms D	Chronos	Energy	D	337	1	366	2
China Air Quality	LOTSA	Climate	H	437	6	397,335	0.3
CMIP6 2000	LOTSA	Climate	6H	8,192	22†	7,300	0.057
ERA5 1989	LOTSA	Climate	H	8,192	15†	8,736	0.085
ERA5 1990	LOTSA	Climate	H	8,192	15†	8,736	0.085
ERA5 1991	LOTSA	Climate	H	8,192	15†	8,736	0.085
Spanish Weather	Kaggle	Climate	H	5	3	24,544	105
Subseasonal	LOTSA	Climate	1D	862	4	16,470	0.3
Subseasonal Precipitation	LOTSA	Climate	1D	862	1	11,323	1.2
Weatherbench daily	Chronos	Climate	1D	10,000†	1	14,610	0.1024
Mexico City Bikes	Chronos	Traffic	H	494	1	104,449	2.5
PEMS04	LOTSA	Traffic	5T	307	3	16,992	1.2
PEMS07	LOTSA	Traffic	5T	883	1	28,224	1.2
PEMS08	LOTSA	Traffic	5T	170	3	17,856	2.1
Q-TRAFFIC	LOTSA	Traffic	15T	45,148	1	5,856	0.024
Taxi (30 Min.)	Chronos	Traffic	30T	2,428	1	1,488	0.88
Taxi (Hourly)	Chronos	Traffic	H	2,428	1	744	0.88
Uber TLC (Hourly)	Chronos	Traffic	H	262	1	4,344	4
Alibaba Cluster Trace 2018	LOTSA	Cloud	5T	58,409	2	1,728	0.009
Wiki Daily	Chronos	Web	D	100,000	1	2,741	0.00512
Monash M3 Monthly	LOTSA	Econ./Fin.	M	1,428	1	126	0.72
NN5 Weekly	LOTSA	Econ./Fin.	W	111	1	105	5
Project Tycho	LOTSA	Health	W	1,258	1	3,854	0.21
Anomaly	TempoPFN	Synthetic	-	5,000	1	10,000	0.0256
ForecastPFN	TempoPFN	Synthetic	-	5,000	1	10,000	1
GP	TempoPFN	Synthetic	-	5,000	1	10,000	0.4096
Kernel Synth 1M	Chronos	Synthetic	-	1,000,000	1	1,024	0.001024
Sawtooth	TempoPFN	Synthetic	-	5,000	1	10,000	0.0512
Sinewave	TempoPFN	Synthetic	-	5,000	1	10,000	0.1024
Spikes	TempoPFN	Synthetic	-	5,000	1	10,000	0.0256
Step	TempoPFN	Synthetic	-	5,000	1	10,000	0.0512
OU	TempoPFN	Synthetic	-	5,000	1	10,000	0.4096

(2) Data Prior, stage 2: synthetic covariate-target dependencies

- **Synthetic multivariate series** built from univariate signals
- A **DAG-based Structural Causal Model**

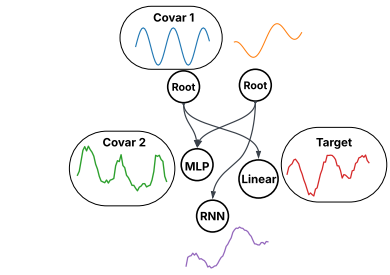


Figure 1: Synthetic target–covariate generation.

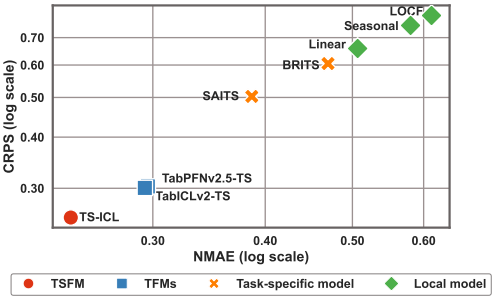
Algorithm 1 Data Prior Sampling Procedure for TS-ICL

Require: Collection of base univariate time series $\mathcal{D}$ (real + synthetic)
Require: SCM registry Υ
Require: Probability π of sampling a univariate task
Sample $u \sim \mathcal{U}(0, 1)$
if $u < \pi$ **then**
  Univariate mode:
  Sample time series $x \sim \mathcal{D}$
  Apply task masking (imputation or forecasting)
  return x
else
  Covariate mode:
  Sample number of task channels:
    $C \sim \text{Exp}(\lambda)$, clipped to $[C_{\min}, C_{\max}]$
  Sample number of root signals:
    $R \sim \text{Geometric}(q)$, clipped to $R_{\max}$
  Define latent DAG size:
    $V = 2C + R$
  Sample root time series:
    $S_1, \ldots, S_R \sim \mathcal{D}$
  Initialize DAG nodes:
    $\{X_i\}_{i=1}^{R} \leftarrow \{S_i\}_{i=1}^{R}$
  DAG construction
  for $i = R + 1, \ldots, V$ **do**
    Sample number of parents $k_i \sim \text{Geometric}(p)$
    Sample parent set $\mathcal{P}_i \subset \{1, \ldots, i-1\}$
    Sample SCM $f_i \sim \Upsilon$
    Compute:
      $X_i \leftarrow \text{Normalize}\big(f_i(\mathbf{X}_{\mathcal{P}_i})\big)$
  end for
  Observation step
  Sample observed set $\mathcal{O} \subset \{1, \ldots, V\}$ such that $|\mathcal{O}| = C$
  Sample target node $y \sim \mathcal{O}$
  Define covariates $\mathbf{x} = \mathcal{O} \setminus \{y\}$
  Task masking
    - imputation: random point/block masking
    - forecasting: future horizon masking
  return $(\mathbf{x}, y)$
end if

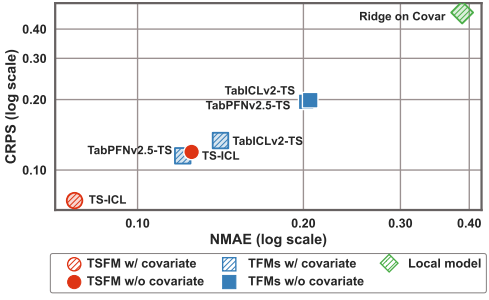
Benchmarks

SOTA zero-shot imputation

TS-ICL improves by $>$ 15% vs TabPFN v2 / TabICL v2 while being $\approx 50\times$ faster at inference



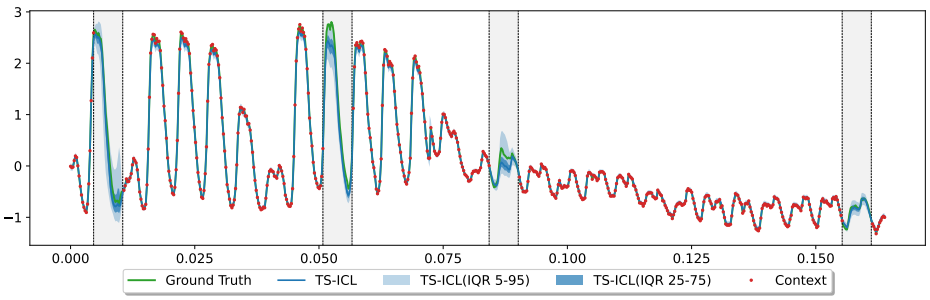
Univariate imputation (132 tasks).



Imputation on 6 tasks with covariates.

fm-impute: 33 OOD datasets - 4 imputation settings / dataset (pointwise / blocks) - 1.3M total windows [12]

SOTA zero-shot imputation



Covid19 Energy - four one-day missing blocks

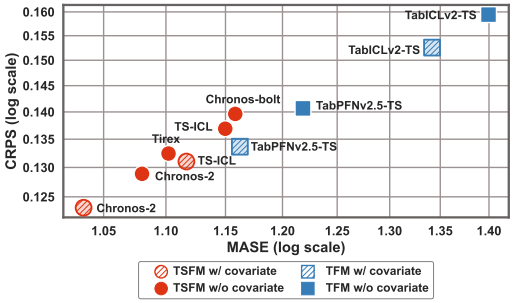
Competitive zero-shot forecasting

TS-ICL outperforms TFMs + 1st generation TSFMs but lags behind recent TSFMs s.a. Chronos-2

Some observations:

- Improvement with informative exogenous covariates
- Smooth forecasts of high-freq phenomenon

↯ Struggles with short contexts



`fev-bench:` *100 tasks, incl. 30 covariate-aware) [14]*

Forecasting with missing values

TS-ICL shows **increased robustness** for forecasting with partially missing historical observations

Table 2: Zero-shot forecasting MASE under increasing missingness in the look-back window for *TS-ICL* and *Chronos-2* on *fev-bench* (100 univariate tasks, look-back = 4092, arithmetic mean). Relative degradation (%) is reported in parentheses. *Seasonal Naive* serves as a simple baseline for evaluating forecasting performance under degraded conditions. Best results are in **bold**.

	0 % missing	30 % missing	50 % missing	70 % missing	90 % missing
Chronos-2	**1.62 (0%)**	2.16 (-33%)	2.44 (-50%)	2.54 (-56%)	3.97 (-144%)
TS-ICL	1.70 (0%)	**1.77 (-4%)**	**1.89** (-11%)	**2.16** (-27%)	3.63 (-113%)
Seasonal(0% missing)	2.48	2.48	2.48	2.48	**2.48**

Conclusion

Take-home messages

TS-ICL: A NEW FAMILY OF FLEXIBLE TSFMS

1. **What is a good TS token?** → TS-ICL as an alternative to patch-based designs - with some cost
2. **SOTA imputation**, **strong forecasting** performance yet still behind leading TSFMs
3. **Synthetic data** will be key for better generalization

Paper

Github

Thanks for listening!

References

[1] Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

[2] Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. Tirex: Zero-shot forecasting across long and short horizons with enhanced in-context learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[3] Ben Cohen, Emaad Khwaja, Youssef Doubli, Salahidine Lemaachi, Chris Lettieri, Charles Masson, Hugo Miccinilli, Elise Ramé, Qiqi Ren, Afshin Rostamizadeh, Jean Ogier du Terrail, Anna-Monica Toon, Kan Wang, Stephan Xie, Zongzhe Xu, Viktoriya Zhukova, David Asker, Ameet Talwalkar, and Othmane Abou-Amal. This time is different: An observability perspective on time series foundation models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[4] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, 2024.

[5] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[6] Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 30583–30598. Curran Associates, Inc., 2022.

[7] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations*, 2022.

[8] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

[9] Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. From tables to time: How TabPFN-v2 outperforms specialized time series forecasting models. *arXiv preprint arXiv:2501.02945*, 2025.

[10] Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024.

[11] Louis Kirsch, James Harrison, Jascha Sohl-Dickstein, and Luke Metz. General-purpose in-context learning by meta-learning transformers. In *Sixth Workshop on Meta-Learning at the Conference on Neural Information Processing Systems*, 2022.

[12] Etienne Le Naour, Tahar Nabil, Adrien Petralia, and Ghislain Agoua. Are time-indexed foundation models the future of time series imputation? *Transactions on Machine Learning Research*, 2026.

[13] Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. *arXiv preprint arXiv:2602.11139*, 2026.

[14] Oleksandr Shchur, Abdul Fatir Ansari, Caner Turkmen, Lorenzo Stella, Nick Erickson, Pablo Guerron, Michael Bohlke-Schneider, and Yuyang Wang. fev-bench: A realistic benchmark for time series forecasting. *arXiv preprint arXiv:2509.26468*, 2025.

[15] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[16] Luisa Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. In *International Conference on Machine Learning*, pages 7693–7702. PMLR, 2019.

Training procedure

We enforce task diversity during training with **randomized settings** for each batch

Imputation

1. Per-batch context length T sampled from:

$$T \sim \begin{cases} \mathcal{U}(128,\,336) & \text{with proba } p_1 = 0.15 \\ \mathcal{U}(512,\,1024) & \text{with proba } p_2 = 0.6 \\ \mathcal{U}(1300,\,1400) & \text{with proba } p_3 = 0.05 \\ \mathcal{U}(2000,\,2100) & \text{with proba } p_4 = 0.1 \\ \mathcal{U}(4000,\,4096) & \text{with proba } p_5 = 0.1 \end{cases},$$

2. **Random Masking Procedure:** A fraction of the obs. are held out as target queries,
   - pointwise rate $\sim \mathcal{U}(\{0.05,\,0.06,\,\ldots,\,0.95\})$;
   - or ≤ 4 blocks of len $\mathcal{U}(12, 168)$

Forecasting

1. (Lookback, horizon) pairs sampled from:

$$(L,\,H) \sim \begin{cases} (\mathcal{U}(50,\,200),\,\mathcal{U}(8,\,20)) & p_1 = 0.2 \\ (\mathcal{U}(200,\,672),\,\mathcal{U}(18,\,36)) & p_2 = 0.15 \\ (\mathcal{U}(1000,\,1100),\,\mathcal{U}(48,\,336)) & p_3 = 0.5 \\ (\mathcal{U}(2000,\,2100),\,\mathcal{U}(48,\,336)) & p_4 = 0.1 \\ (\mathcal{U}(3062,\,4096),\,\mathcal{U}(48,\,672)) & p_5 = 0.05 \end{cases}.$$

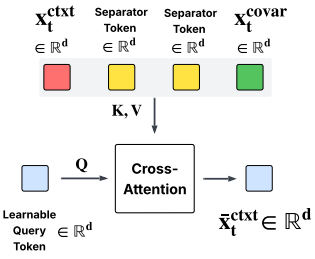
2. **Random Masking Procedure:** A fraction $\rho \sim \mathcal{U}(\{0.1,\,0.25,\,0.5,\,0.75\})$ of the obs. in the lookback window are removed with probability 0.15.

Training hyperparameters

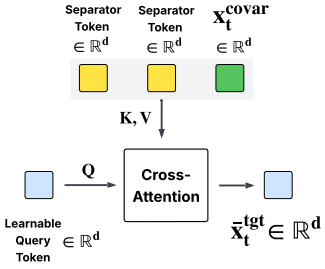
- **Base architecture:** 28M parameters
- **Loss:** 99-quantile regression (pinball loss) $\rightarrow$ *probabilistic model*
- **Optimizer:** *Muon* [10] for 2D parameters and *AdamW* for 1D parameters
- **Learning rate:** max 4×10^{-4}
- **Scheduler:** cosine decay down to 5×10^{-4}
- **Hardware:** $4\times$ NVidia H100 GPUs (92GB)
- **Batch size:** global batch size $B = 256$ (mini-batch size: 32 + grad. accumulation: 2 steps)
- **Training budget:**
  1. *Imputation checkpoint:* 500k optimization steps, $\approx$ 5 training days
  2. *Forecasting checkpoint:* 650k optimization steps, $\approx$ 9 training days

Covariate: current bottleneck

Covariates are fed as inputs to the ICL Transformer: a *learnable label token* merges information from the covariate(s) and the target. 2 ”separator” tokens differentiate target vs covariate.



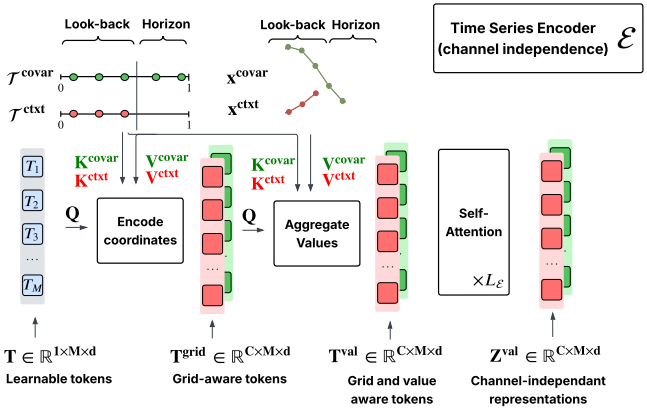
*Building the **context** labels*



*Building the **target** labels*

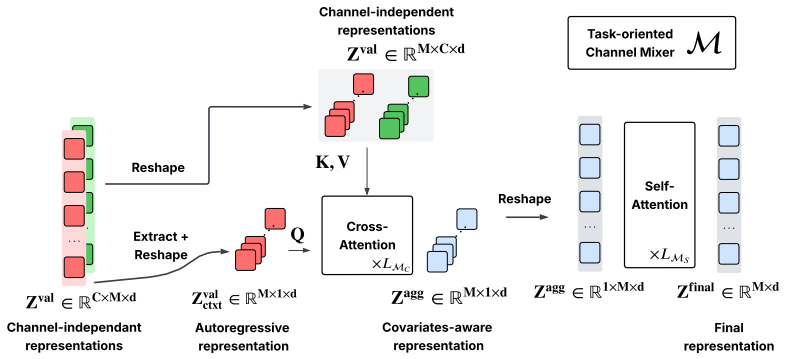
(i) Time Series Encoder - with channel independence

Goal: map the observed context and covariates jointly into a channel-indep. latent representation



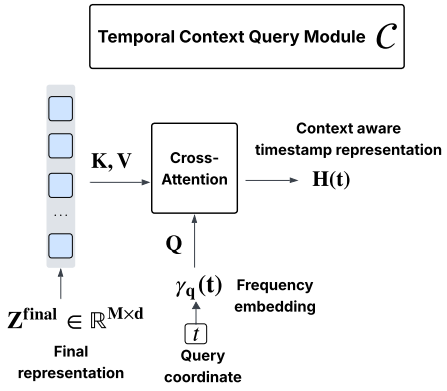
(ii) Channel Mixer

Goal: aggregate information across multiple channels



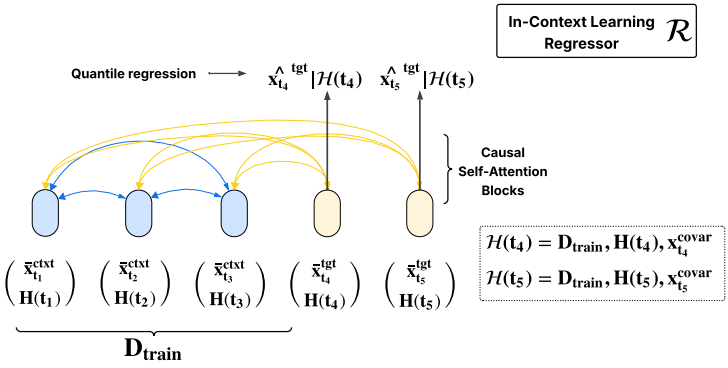
(iii) Temporal Context Query module

Goal: bridge discrete tokens w. time domain → continuous-time querying of the encoded context



(iv) In-Context Learning Regressor

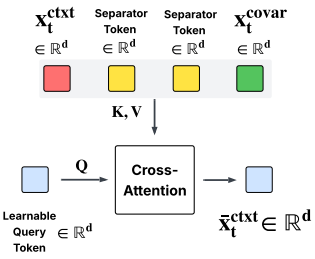
Goal: solve the in-context regression task



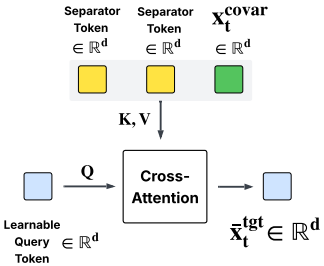
(iv) In-Context Learning Regressor

Goal: solve the in-context regression task

Regression labels obtained via x-attn to accomodate any covar. setting (+ missing target values)



*Building the **context** labels*



*Building the **target** labels*